Caracterización del Espacio Web de Perú

Gabriel H. Tolosa, Fernando R. A. Bordignon y Pablo J. Lavallén

Universidad Nacional de Luján, Argentina
Departamento de Ciencias Básicas
Laboratorio de Redes
{tolosoft, bordi, plavallen}@unlu.edu.ar

Resumen

La web se presenta como un espacio público utilizado por múltiples usuarios con objetivos diferentes. Originalmente, se presentaba como un repositorio distribuido que permitía compartir información y - aunque no ha perdido este objetivo - en la actualidad es un medio de publicación y servicio para diferentes usos como comercio, publicidad, educación, entretenimiento y contactos sociales, entre otros.

Si bien la web se encuentra en constante crecimiento el estudio de características y tendencias entrega una valiosa información, tanto para entender su estructura como para desarrollar herramientas que faciliten la utilización de sus recursos. Este trabajo tiene por objetivo principal la caracterización del espacio web de Perú, en el marco de las nuevas tendencias de crecimiento y evolución. Se presentan características de los sitios, de las páginas, tecnologías utilizadas y aspectos descriptivos de su topología.

1 – Introducción

En la actualidad la World Wide Web (web) se presenta como un espacio público utilizado por múltiples usuarios con objetivos diferentes. Originalmente, se planificó como un repositorio distribuido para compartir información y – aunque no ha perdido este objetivo – hoy es un medio de publicación y servicios para diferentes usos, tales como educación, comercio, publicidad, entretenimiento y contactos sociales, entre otros.

La web propone otras dimensiones que la diferencian de bases de datos y bibliotecas tradicionales [Björneborn, 2001] ya que sobre este espacio interactúan millones de actores quienes dinámicamente agregan, quitan y modifican páginas y enlaces, es decir, su contenido y su estructura presentan características propias que no poseen otros repositorios de información.

La cantidad y diversidad de sus usuarios y la variedad de sus recursos genera que su estado se tome como un indicador del desarrollo tecnológico de una comunidad, en particular, en el contexto del concepto de “Sociedad del la Información”, entendida como la capacidad de sus miembros (individuos, organizaciones, gobiernos) para acceder y compartir información desde cualquier lugar, por medios electrónicos. En esta dirección, la UNESCO reconoce que el desarrollo de Internet y las tecnologías digitales son clave su evolución [UNESCO, 2005].

De aquí que se plantea la necesidad de caracterizar su estructura y contenido a los efectos de desarrollar modelos, técnicas y herramientas que faciliten el acceso y utilización de sus recursos. Dado su tamaño, el cual es virtualmente infinito, se presenta como un desafío en la gestión eficiente de sus recursos.

La caracterización de espacios web es una tarea compleja a escala global por lo que se han realizado estudios sobre dominios nacionales [Baeza-Yates, 2003] [Baeza-Yates, 2005_a] [Baeza-Yates, 2005_b] [Efthimiadis, 2004] [Gomes, 2005] [Meneses, 2006] [Modesto, 2005]. En éstos, se extraen muestras del conjunto de interés utilizando diversas técnicas y se estudian algunos aspectos de estructura y contenidos existentes. Los análisis estructurales ofrecen – además – indicadores de accesibilidad de la información contenida en sitios y páginas [Amat, 2003].

Como aporte de esta investigación se presentan algunos resultados sobre un trabajo de caracterización del espacio web de Perú, a los efectos de obtener una fotografía inicial sobre la cual identificar fortalezas, debilidades y oportunidades de crecimiento y evolución. Al conocer su estado de desarrollo se cuenta con una base para planificar acciones tendientes a expandir su tamaño y mejorar su calidad.

Como aporte complementario, se generó una página web con un modelo de visualización de relaciones entre sitios. Donde a partir de ingresar una dirección de sitio se retorna un applet que despliega un gráfico interactivo con todos los sitios relacionados, ya sea por enlaces entrantes o salientes. La aplicación se encuentra disponible en http://www.tyr.unlu.edu.ar/research/webpe/.

2 – Recolección de Datos

La recolección de la muestra se realizó con el software *crawler* WIRE [Castillo, 2005] en el mes de agosto del año 2006. Como infraestructura informática, se utilizó un equipo servidor con un procesador de 2.4 GHz y 1 GB de memoria RAM corriendo el sistema operativo Linux Debian Sarge y un enlace a Internet de 2 Mbps.

El módulo de recolección se configuró para descargar únicamente páginas web bajo el dominio ".pe". El *crawler* fue inicialmente alimentado con más de 15.000 direcciones de dominios provistos por el NIC Perú[1]. Al módulo de *crawling* se lo configuró con los siguientes parámetros: profundidad máxima en páginas dinámicas: 5 y profundidad máxima en páginas estáticas: 15. Por otro lado, se limitó la recolección sólo a páginas HTML hasta un máximo de 100 KB y hasta 15.000 páginas por sitio.

El proceso de recolección finalizó cuando en *crawler* no encontró más URLs bajo el dominio de estudio, por los que la muestra obtenida es interesante ya que corresponde al conjunto de páginas más visibles o mejor conectadas.

[1] http://www.nic.pe/

3 – Características Generales

Se descargaron 1.629.745 páginas desde 8.908 sitios, que corresponden a 7.945 dominios de tercer nivel. Los datos brindados se consideran como una cota inferior en el tamaño de la web de Perú, dado que pueden existir sitios que aún no se encuentran conectados con esta porción.

Orden	Dominio	Cantidad de Sitios
1	perucultural.org.pe	69
2	uni.edu.pe	64
3	unitru.edu.pe	48
4	pucp.edu.pe	48
5	terra.com.pe	36
6	rcp.net.pe	32
7	unmsm.edu.pe	27
8	usmp.edu.pe	23
9	upeu.edu.pe	19
10	upc.edu.pe	16

Tabla 1: Distribución de dominios con mayor cantidad de sitios (primeros diez).

El 93,32% de las páginas son únicas y el 6,68% se encuentran duplicadas. El 64,47% son páginas estáticas, mientras que las dinámicas suman el 35,53%. A partir de datos aportados por NIC Perú se determinó que proporción de sitios están actualmente activos, la información se agrupa por dominios de segundo nivel:

	Registrados en NIC Perú (agosto 2006)		Dominios activos[2]		Proporción entre registrados y activos
Dominio de 2do nivel	Cantidad	Fracción	Cantidad	Fracción	
com.pe	12.942	0,8016	5.844	0,7343	0,45
org.pe	1.335	0,0827	863	0,1084	0,65
edu.pe	915	0,0567	683	0,0858	0,75
gob.pe	858	0,0531	532	0,0668	0,62
net.pe	70	0,0043	19	0,0024	0,27
mil.pe	13	0,0008	9	0,0011	0,69
sld.pe	8	0,0005	5	0,0006	0,63
nom.pe	3	0,0002	3	0,0004	1,00
nic.pe	2	0,0001	1	0,0001	0,50
Total	16.146	1	7.959	1	0,49

Tabla 2: Distribución de dominios de segundo nivel

La distribución de los principales códigos de estado entregados por los servidores web al *crawler* al intentar descargar cada recurso se presentan en la tabla 3. Dichos códigos fueron divididos en las siguientes categorías:

[2] Se considera dominios activos a aquellos se descargo exitosamente al menos un documento

Categoría		%
OK	Peticiones con estado de éxito	89,36
MOVED	Códigos que indican que el servidor redirige la petición a otra URL alternativa	4,19
SERVER ERROR	Peticiones que arrojan por resultado una falla producida del lado del servidor	0,93
FORBIDDEN	Peticiones que son denegadas por el servidor.	0,34
NOT FOUND	Recurso inexistente	5,18

Tabla 3: Distribución de códigos de estado

Se observa un 89,36% de transferencias exitosas lo cual demuestra que no hubo problemas mayores de conectividad, ni de disponibilidad de servicios.

4 – Sitios y Páginas

En lo relativo a sitios, se obtuvieron datos sobre cantidad promedio, edad y enlaces de sus páginas. Tales observaciones – entre otras – se presentan en la tabla 4.

Sitios que aportaron recursos	8.908
Sitios sin enlaces entrantes	5.688
Sitios sin enlaces salientes	5.948
Sitios con edad correcta en páginas	7.464
Promedio de páginas por sitio	179,85
Promedio de páginas estáticas por sitio	116,23
Promedio de dinámicas por sitio	73,52
Promedio de edad de la página más antigua (meses)	14,93
Promedio de edad de la página más nueva (meses)	8,35
Promedio del tamaño de sitios (en MB)	2,7
Promedio de profundidad de sitios	3,12

Tabla 4 – Información resumen de sitios

4.1 – Tamaño de los sitios

Se presenta el tamaño de los sitios en dos formas: cantidad de información que poseen (tamaño del sitio en bytes) y cantidad de documentos. En la tabla 5 se muestran los primeros quince sitios de mayor volumen.

Orden	Sitios por tamaño	Sitios por cantidad de documentos
1	www.booking.com.pe	foros.hispavista.com.pe
2	*www.mininter.gob.pe*	*www.socialista.org.pe*
3	directorio.internet.com.pe	*www.mininter.gob.pe*
4	*www.scout.org.pe*	www.internet.com.pe
5	sexualidadsana.com.pe	directorio.internet.com.pe

6	*www.sni.org.pe*	sexualidadsana.com.pe
7	foros.hispavista.com.pe	*www.bibliocentral.udep.edu.pe*
8	forum.sugoi.com.pe	www.yaclasificados.com.pe
9	www.yaclasificados.com.pe	*www.sni.org.pe*
10	*www.bibliocentral.udep.edu.pe*	marriott.com.pe
11	download.terra.com.pe	*cies.org.pe*
12	zdnet.terra.com.pe	www.adobe.com.pe
13	www.internet.com.pe	*blog.pucp.edu.pe*
14	salud.terra.com.pe	salud.terra.com.pe
15	www.iwantu.com.pe	download.terra.com.pe

Tabla 5 – Primeros quince sitios con mayor cantidad de contenidos

Las distribuciones de tamaño y cantidad de sitios siguen leyes de potencia ya que existen pocos sitios que poseen valores altos y hay muchos con valores bajos [Adamic, 2002]. Una observación interesante es que existen en estas primeras ubicaciones sitios que no pertenecen al ámbito comercial (desatacados en itálicas).

4.2 – Páginas por Sitio

En esta sección se presentan resultados sobre la distribución de páginas por sitio. Como se puede observar existe más del 55% de sitios con un máximo de 10 páginas, lo que indica que esta parte del espacio web estudiado se encuentra muy poco desarrollada. Complementariamente, existen muy poco sitio grandes (>5000 páginas). La distribución de esta variable se presenta en la tabla 6.

Rango	Sitios	%
Entre 1 y 10 páginas	5.040	56,58
Entre 11 y 50 páginas	2.125	23,85
Entre 51 y 500 páginas	1.368	15,36
Entre 501 y 5000 páginas	300	3,37
Más de 5000 páginas	75	0,84

Tabla 6 – Distribución de la cantidad de páginas por sitio

El promedio de páginas por sitio (179) es más alto que en observaciones realizadas en otros estudios, por ejemplo, en Chile se encontraron 58 [Baeza-Yates, 2005a] y 66 en Brasil [Modesto, 2005] pero mucho menor a lo reportado para Costa Rica (514 páginas) [Meneces, 2006].

4.3 – Enlaces entre Sitios

Otra cuestión interesante se observa en el estudio de enlaces entre sitios dentro del dominio. Aquí cabe mencionar que los enlaces son indicadores importantes de la popularidad de los sitios web, especialmente los enlaces entrantes ya que son utilizados por diferentes algoritmos de rankings [Page, 1998] [Kleinberg, 1999]. Hallamos 5.688 sitios (64%) sin enlaces entrantes y 5.948 (68%) sin enlaces salientes. En el primero de los casos se puede comentar que estos sitios tienen problemas de ranking por los motores de búsqueda principales mientras que en el segundo, existe un porcentaje elevado de sitios que no hacen referencia al espacio web de su propio país.

4.4 – Contenido de las Páginas

A los efectos de brindar una pequeña caracterización del contenido de las páginas, construimos una nube de etiquetas (Tag-Cloud) [Godwin-Jones, 2006] con aquellos términos más representativos de acuerdo a su frecuencia de aparición (Figura 1).

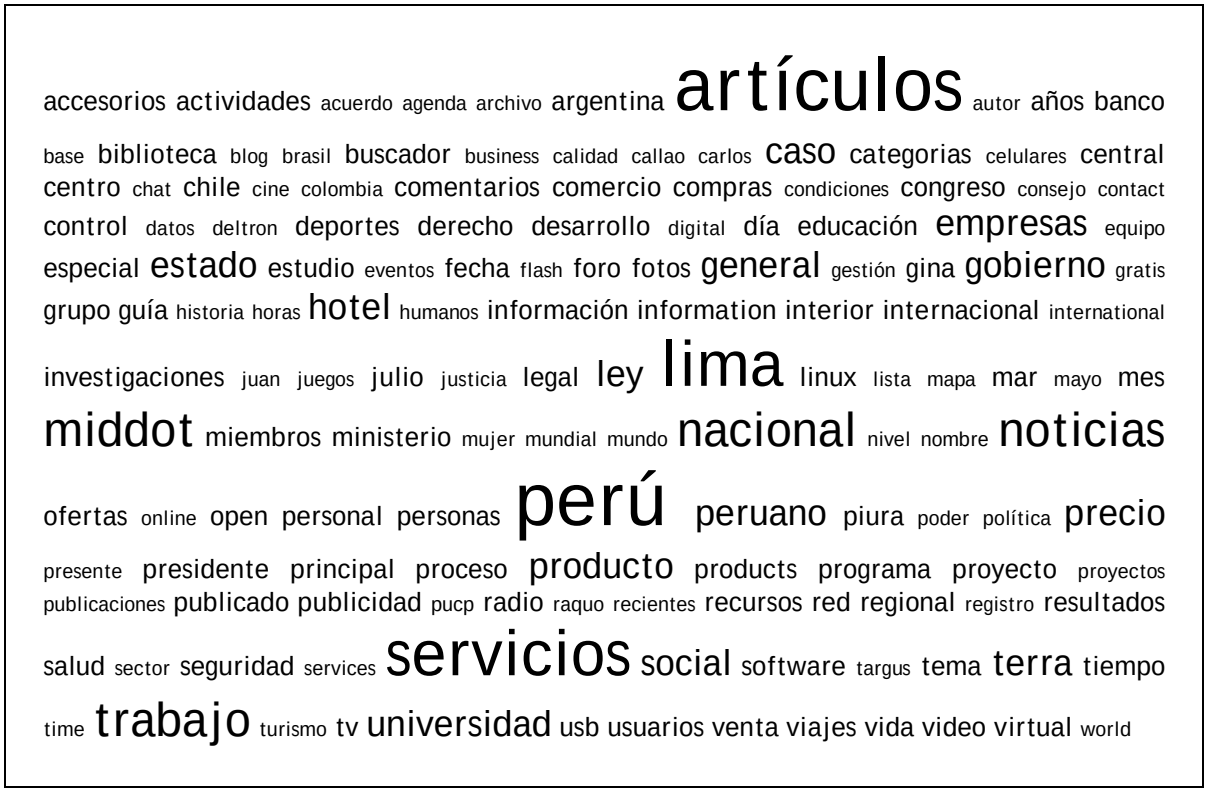

Figura 1 – Nube de etiquetas basada en el contenido de las páginas web

Aquí se puede apreciar de forma gráfica los términos más representativos (mayor tamaño de letra) y – si bien se utilizan generalmente como una alternativa de navegación – presenta una vista distinta del contenido de todo un espacio web.

5 – Enlaces y Rankings

También se estudió el espacio web de Perú como grafo dirigido a nivel de páginas (webgraph), donde éstas son los nodos y los enlaces las aristas [Broder, 2000].

5.1 – Distribuciones de Grado

Las variables analizadas aquí corresponden al grado entrante y saliente a nivel de páginas. El grado entrante (*in degree*) es el número de enlaces que apuntan a una página y es un indicador de reputación utilizado para armar rankings de salida por algunos motores de consulta. Por otro lado, el grado saliente (*out degree*) es el número de enlaces salientes que posee una página. En ambos casos, las distribuciones de estos valores se ajustan a leyes de potencias como lo hallado en

otras muestras de la web. En la tabla 7 se resumen los resultados.

In degree Cantidad de páginas	%	Rango	Out degree Cantidad de páginas	%
557029	34,18	0	972446	59,67
918026	56,33	de 1 a 5	191841	11,77
62862	3,86	de 6 a 10	111888	6,87
37037	2,27	de 11 a 20	171897	10,55
41848	2,57	de 21 a 100	172235	10,57
11816	0,73	de 101 a 1000	9438	0,58
1127	0,07	de 1001 en adelante		

Tabla 7 – Distribuciones de grado

Nótese un alto porcentaje de páginas con grado entrante 0, las cuales tienen claramente posibles problemas de visibilidad y rankeo. De forma similar, hay casi un 60% de páginas con grado saliente 0, lo que si bien no genera el inconveniente mencionado, es un indicador de baja conectividad para este espacio web.

5.2 – Rankings de Sitios

A continuación (Tabla 8) se presentan resultados acerca de los 15 mejores sitios rankeados utilizando diferentes algoritmos: Hub y Authorities, del algoritmo HITS [Kleinberg, 1999] y Siterank, basado en el algoritmo PageRank [Page, 1998].

Orden	Ranking: Hub	Ranking: Authority	Ranking: Siterank
1	directorio.internet.com.pe	terra.com.pe	sexualidadsana.com.pe
2	www.internet.com.pe	www.larepublica.com.pe	www.yaclasificados.com.pe
3	*www.mininter.gob.pe*	www.deltron.com.pe	*www.bibliocentral.udep.edu.pe*
4	tiempo.terra.com.pe	*www.cverdad.org.pe*	*cies.org.pe*
5	mundial2002.terra.com.pe	*www.upc.edu.pe*	*www.sni.org.pe*
6	www.adoos.com.pe	*www.pucp.edu.pe*	*www.embajadasuiza.org.pe*
7	*info.upc.edu.pe*	*www.unmsm.edu.pe*	www.trabajo.com.pe
8	*dia.pucp.edu.pe*	*www.sat.gob.pe*	*www.scout.org.pe*
9	www.losclasificados.com.pe	www.expreso.com.pe	www.subastas123.com.pe
10	speedy3.deltron.com.pe	www.gestion.com.pe	*www.unmsm.edu.pe*
11	*www.socialista.org.pe*	www.libero.com.pe	www.booking.com.pe
12	*blog.pucp.edu.pe*	*www.sunat.gob.pe*	*www.cepis.org.pe*
13	w2-bs.deltron.com.pe	*mintra.gob.pe*	*www2.congreso.gob.pe*
14	www-spd.deltron.com.pe	larazon.com.pe	*www.cverdad.org.pe*
15	*www.ucsp.edu.pe*	*www.inei.gob.pe*	www.computrabajo.com.pe

Tabla 8 – Rankings de hubs, autoridades y siterank

Aquí – nuevamente – destacamos los sitios no comerciales (itálicas), los cuales se encuentran en una proporción interesante. Esto brinda la idea que la web de Perú no se encuentra “dominada” por sitios comerciales que cubran los primeros puestos de los rankings.

5.3 – Macroestructura del Espacio Web

Se analizó la composición del espacio web de Perú de acuerdo a la metodología de Broder [Broder, 2000] que plantea 6 regiones en las cuales ubicar a las páginas, a saber:

MAIN:	Componente fuertemente conexa principal
MAIN-MAIN:	Sitios relacionados directamente con IN y con OUT.
MAIN-IN:	Sitios relacionados directamente con IN, pero no con OUT.
MAIN-OUT:	Sitios relacionados directamente con OUT, pero no con IN.
MAIN-NORM:	Sitios en MAIN que no corresponden a ninguna de las categorías vistas.
IN:	Sitios que llegan a MAIN, pero de MAIN no se puede llegar a ellos
OUT:	Sitios a los que se llega desde MAIN, pero no se puede retornar
TUNNEL:	Sitios en caminos de IN a OUT sin atravesar MAIN.
TENTACLE:	Sitios a los que se llega de IN o van a OUT, y no están en MAIN ni en TUNNEL.
ISLANDS:	Sitios no conectados a nada de lo anterior

Como se aprecia en la tabla 9, el espacio web de Perú se halla débilmente interconectado ya que la componente MAIN es pequeña y existe un 53% de sitios en la región islas. Esta es una indicación más de la necesidad de mejorar la estructura de vínculos para permitir una mejor integración, mejores rankings y mayor visibilidad a nivel sitios.

Componente	Sitios	Fracción
MAIN	1.334	0,15
MAIN NORM	*368*	
MAIN MAIN	*380*	
MAIN IN	*284*	
MAIN OUT	*302*	
IN	1.129	0,13
OUT	1.283	0,14
Tentacle IN	254	0,03
Tentacle OUT	150	0,02
Tunnel	12	0,00
Islands	4.746	0,53

Tabla 9 – Macroestructura del espacio web de Perú

6 – Conclusiones

El espacio web de la República del Perú tiene al menos 8.908 sitios activos que corresponden a 7.945 dominios. De ellos, en este trabajo, se recolectaron 1.629.745 páginas web. El 49% de los dominios registrados en NIC Perú están activos, es decir que al menos existe un servidor web respondiendo peticiones.

Existe cerca de un 5,8% de enlaces rotos. El promedio de documentos por sitio es de alrededor de 180 y su tamaño promedio es de 2,7 MB. El promedio de profundidad de los sitios es de 3,12 niveles.

Se puede observar una interesante participación de organizaciones no comerciales ya que tanto en tamaño de sitios como en los rankings hay aproximadamente un 50% que pertenecen al gobierno, la educación y otras entidades sin fines de lucro.

En cuanto a los enlaces (estudiado a nivel de páginas y sitios) se evidencia una conectividad deficiente. Hay un 64% de sitios sin enlaces entrantes y un 68% sin enlaces salientes. Además, la componente fuertemente conectada (MAIN) posee solo el 15% de los sitios, mientras que las islas suman el 53%.

7 – Referencias

[Amat, 2003] C. B. Amat. Caracterización de una muestra de sedes Web españolas bajo dominio .es. Boletín de Red IRIS, Nro 64, pp 33-40, 2003.

[Baeza-Yates, 2003] R. Baeza Yates, B. Poblete y F. Saint-Jean, Evolución de la Web Chilena 2001-2002, Estudio Técnico, Centro de Investigación de la Web (CIW), 2003.

[Baeza-Yates, 2005_a] R. Baeza-Yates and C. Castillo. Características de la Web Chilena 2004. Technical Report, Center for Web Research, University of Chile, 2005.

[Baeza-Yates, 2005_b] R. Baeza-Yates, C. Castillo and V. Lopez. Characteristics of the Web of Spain. Cybermetrics, Vol. 9, No. 1, 2005.

[Björneborn, 2001] L. Björneborn y P. Ingwersen. Perspectives of Webometrics. Scientometrics, Vol. 50, Nro. 1 pp 65-82, 2001.

[Broder, 2000] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, A. y J. Wiener. Graph Structure in the Web. Proc. 9th International World Wide Web Conference (WWW9)/Computer Networks, 33(1-6),2000, pp. 309-320.

[Castillo, 2005] C. Castillo and R. Baeza-Yates. WIRE: an Open Source Web Information Retrieval Environment. Workshop on Open Source Web Information Retrieval (OSWIR), 2005.

[Efthimiadis, 2004] E. Efthimiadis and C. Castillo. Charting the Greek Web. In Proceedings of the Conference of the American Society for Information Science and Technology (ASIST), Providence, Rhode Island, USA, November, 2004.

[Godwin-Jones, 2006] R. Godwin-Jones. EMERGING TECHNOLOGIES Tag Clouds in the Blogosphere: Electronic Literacy and social Networking. Language Learning & Technolog, Vol. 10, No. 2, pp. 8-15, 2006.

[Gomes, 2005] D. Gomes and M.J. Silva. Characterizing a National Community Web.
ACM Transactions on Internet Technology, 5(3):508–531, 2005.

[Kleinberg, 1999] J. Kleinberg. Authoritative Sources in a Hyperlinked Environment. Journal of the ACM, 46(5):604-632, 1999.

[Meneses, 2006] E. Meneses Mining the Costa Rican Web. International Conference on Web Information Systems and Technologies. Setúbal, Portugal. Abril, 2006.

[Modesto, 2005] M. Modesto, A. Pereira, N. Ziviani, C. Castillo and R. Baeza-Yates. Un Novo Retrato da Web Brasileira. In Proceedings of SEMISH, São Leopoldo, Brazil, 2005.

[Page, 1998] L. Page, S. Brin, R. Motwani y T. Winograd. The PageRank citation ranking: bringing order to the Web. Technical report, Stanford Digital Library Technologies Project, 1998.

[UNESCO, 2005] Informe Mundial de la UNESCO: Hacia las Sociedades del Conocimiento, 2005.